АППАРАТНО-ПРОГРАММНАЯ ИНФРАСТРУКТУРА ДЛЯ РЕАЛИЗАЦИИ УЧЕБНЫХ ПРОЕКТОВ ПО ТЕМАТИКЕ ОБРАБОТКИ ПОТОКОВ ДАННЫХ В ВЫСОКОНАГРУЖЕННЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

В.И. Проценко, П.Г. Серафимович, С.Б. Попов, Н.Л. Казанский

Институт систем обработки изображений РАН, Самара, Россия,
Самарский государственный аэрокосмический университет имени академика С.П. Королёва (национальный исследовательский университет) (СГАУ), Самара, Россия

В работе анализируется современное состояние аппаратно-программных средств обработки интенсивных потоков данных в реальном времени. Рассмотрены проблемы построения высокопроизводительного вычислительного комплекса обработки потоков данных на основе технологии IBM BigInsights. Демонстрируются возможности программного пакета Apache Spark для обработки временных рядов. Приведены примеры получения некоторых статистических характеристик временных рядов в реальном времени.

Ключевые слова: обработка потока данных, высоконагруженные информационные системы, статистика временных рядов.

IBM InfoSphere BigInsights предоставляет богатый набор аналитических возможностей, позволяющий предприятиям выполнять низкозатратный анализ массивных неструктурированных и структурированных данных в их исходном формате. Сочетает программное обеспечение с открытым кодом Apache Hadoop с инновациями IBM, включая сложную текстовую аналитику, IBM BigSheets для анализа данных и разнообразные функции для повышения производительности, безопасности и администрирования. Результатом является низкозатратное решение для сложной аналитики больших объемов данных с дружественным интерфейсом.

Apache Spark (от англ. spark — искра, вспышка) — программный каркас с открытым исходным кодом для реализации распределённой обработки неструктурированных и слабоструктурированных данных, входящий в экосистему проектов Hadoop. В отличие от классического обработчика из ядра Hadoop, реализующего двухуровневую концепцию MapReduce с дисковым хранилищем, данная технология использует специализированные примитивы для рекуррентной обработки в оперативной памяти, благодаря чему позволяет получать значительный выигрыш в скорости работы для некоторых классов задач [2], в частности, возможность многократного доступа к загруженным в память пользовательским данным делает библиотеку привлекательной для алгоритмов машинного обучения[3].

Проект предоставляет программные интерфейсы для языков Java, Scala, Python, R. Написан в основном на Scala. Состоит из ядра и нескольких расширений, таких как Spark SQL (позволяет выполнять SQL-запросы над данными), Spark Streaming (надстройка для обработки потоковых данных) [1-8], Spark MLib (набор библиотек машинного обучения), GraphX (предназначено для распределённой обработки графов). Может работать как в среде кластера Hadoop под управлением YARN, так и без компонентов ядра Hadoop, поддерживает несколько распределённых систем хранения — HDFS, OpenStack Swift, NoSQL-СУБД Cassandra, Amazon S3.

Необходимость обработки временных рядов определило появление множества математико-статистических методов анализа, предназначенных для выявления структуры временных рядов и для их прогнозирования. Сюда относятся, в частности, методы регрессионного анализа. Выявление структуры временного ряда необходимо для того, чтобы построить математическую модель того явления, которое является источником анализируемого временного ряда. Прогноз будущих значений временного ряда используется для эффективного принятия решений.

Литература

1. Twardowski, B. Multi-agent Architecture for Real-Time Big Data Processing / Bartlomiej Twardowski and Dominik Ryzko // Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on. – 2014. –Vol. 3. – P. 333-337.
2. Osman, A. Towards real-time analytics in the cloud. / Ahmed Osman, Mohamed El-Refaey and Ayman Elnaggar // Services (SERVICES), 2013 IEEE Ninth World Congress on. – 2013. – P. 428-435.
3. Shoro, A.G. Big Data Analysis: Apache Spark Perspective / Abdul Ghaffar Shoro and Tariq Rahim Soomro // Global Journal of Computer Science and Technology. – 2015. – Vol. 15(1). – P. 7-14.
4. Namiot, D. On Big Data Stream Processing / Dmitry Namiot // International Journal of Open Information Technologies. – 2015. – Vol. 3(8). – P. 48-51.
5. Gradvohl, A.L.S. Comparing distributed online stream processing systems considering fault tolerance issues / André Leon Sampaio Gradvohl, Hermes Senger, Luciana Arantes and Pierre Sens // Journal of Emerging Technologies in Web Intelligence. – 2014. – Vol. 6(2). – P. 174-179.
6. Stonebraker, M. The 8 requirements of real-time stream processing / Michael Stonebraker, Uğur Çetintemel and Stan Zdonik // ACM SIGMOD Record. – 2005. – Vol. 34(4). – P. 42-47.
7. Papadimitriou, S. Streaming pattern discovery in multiple time-series / Spiros Papadimitriou, Jimeng Sun and Christos Faloutsos // Proceedings of the 31st international conference on Very large data bases, VLDB Endowment. – 2005. – P. 697-708.
8. Data streams: models and algorithms / ed. Charu C. Aggarwal. – Springer Science & Business Media, LLC.: 2007.